Predictable patterns of CTL escape and reversion across host populations and viral subtypes in HIV-1 evolution

Duncan S. Palmer[*1,2,3], Emily Adland[4], John Frater[2,5,6], Philip J. R. Goulder[4,7], Thumbi Ndung'u[7], Philippa C. Matthews[5], Rodney E. Phillips[2,5], Roger Shapiro[8,9], Gil McVean[1,3], and Angela R. McLean[2,10]

[1]Department of Statistics, 1 South Parks Road, University of Oxford, Oxford OX1 3TG, UK
[2]Institute for Emerging Infections, The Oxford Martin School, Oxford OX1 3BD, UK
[3]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK
[4]Department of Paediatrics, University of Oxford, Oxford OX1 3SY, UK
[5]Nuffield Department of Clinical Medicine, Peter Medawar Building for Pathogen Research, University of Oxford, Oxford OX1 3SY, UK
[6]Oxford NIHR Biomedical Research Centre, Oxford OX3 7LE, UK
[7]HIV Pathogenesis Programme, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, 4013 South Africa
[8]Botswana Harvard AIDS Institute Partnership, Gaborone BO 320, Botswana
[9]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02215
[10]Zoology Department, South Parks Road, University of Oxford, Oxford OX1 3PS, UK

Abstract

The twin processes of viral evolutionary escape and reversion in response to host immune pressure, in particular the cytotoxic T-lymphocyte (CTL) response, shape Human Immunodeficiency Virus-1 sequence evolution in infected host populations. The tempo of CTL escape and reversion is known to differ between CTL escape variants in a given host population. Here, we ask: are rates of escape and reversion comparable across infected host populations? For three cohorts taken from three continents, we estimate escape and reversion rates at 23 escape sites in optimally defined Gag epitopes. We find consistent escape rate estimates across the examined cohorts. Reversion rates are also consistent between a Canadian and South African infected host population. Certain Gag escape variants that incur a large replicative fitness cost are known to revert rapidly upon transmission. However, the relationship between escape/reversion rates and viral replicative capacity across a large number of epitopes has not been interrogated. We investigate this relationship by examining *in vitro* replicative capacities of viral sequences with minimal variation: point escape mutants induced in a lab strain. Remarkably, despite the complexities of epistatic effects exemplified by pathways to escape in famous epitopes, and the diversity of both hosts and viruses, CTL escape mutants which escape rapidly tend to be those with the highest replicative capacity when applied as a single point mutation. Similarly, mutants inducing the greatest costs to viral replicative capacity tend to revert more quickly. These data suggest that escape rates in *gag* are consistent across host populations, and that in general these rates are dominated by site specific effects upon viral replicative capacity.

Keywords: HIV-1, escape, CTL, phylodynamics, epistasis

Abbreviations: HIV-1, Human Immunodeficiency Virus-1; CTL, Cytotoxic T-lymphocyte, MCMC; Markov Chain Monte-Carlo

In HIV-1 infected individuals there is a constant arms race between the virus on one side and the host immune response on the other [1–3]. The cytotoxic T-lymphocyte (CTL) response - a part of the adaptive immune system - is a major driving force of HIV evolution both within and across infected individuals [4, 5]. Within nucleated cells, cytosolic peptides (both self and non-self) are presented at

---

[*]duncan.stuart.palmer@gmail.com

the cell surface by human leukocyte antigen (HLA) class I molecules [6], encoded by HLA genes; the most variable loci in the entire human genome [7]. These protein fragments are known as 'epitopes' upon presentation. CTLs may then recognise epitopes as non-self and destroy the presenting cell. The CTL response can select for mutations within or flanking epitopes which result in reduced immune recognition of virally infected cells. These variants, termed CTL escape mutants, may reduce HLA class I binding affinity, alter epitope processing, or affect T-cell receptor contact sites [8–11]. Importantly, the repertoire of viral epitopes that may be presented is dependent upon the 'type' of the HLA class I molecule. If an epitope can be presented, it is said to be 'restricted' by that HLA type, and the host is known as 'HLA matched' for that epitope. A host lacking the class I molecule required to present the epitope is known as 'HLA mismatched' for that epitope. Viruses bearing escape mutations can be transmitted between hosts [12, 13], and reversion of the infecting virus may take place due to the removal of HLA dependent selection pressure and viral fitness costs of CTL escape mutants [14–17].

The viral population present in a given infected individual reflects the selective environment of the current host as well as the remnants of selection from previous hosts. During the course of HIV-1 infection, viral mutations can open pathways to CTL immune escape or compensate for costs to replicative capacity of such mutations [18–22], which can lead to a delay in reversion upon subsequent transmission [20]. Methods used to model and parameterise the CTL escape and subsequent reversion often assume that the timings of escape and reversion are governed by behaviour at a single site [23–30]. This is a clear simplification. The existence and potential complexity of epistatic interactions, coupled with the diversity in HLA profiles, CTL killing efficiency, drug regimes, and vertical T-cell immunodominance [31] leads to the following questions: can individual escape mutants be characterised as having their own escape and reversion rates in a given infected host population, and are escape and reversion rates consistent across host populations?

The ideal data to answer these questions would consist of longitudinal viral sequence samples starting early in infection from a large number of hosts with known HLA types. However, this is not feasible. Instead, we have developed a method which allows us to use cross-sectional viral sequence data and host HLA information to estimate escape and reversion rates whilst accounting for the underlying viral genealogy [24]. This allows us to estimate population specific rates of escape and reversion at escape sites within optimally defined [32] epitopes. By comparing the resultant escape/reversion rate estimates, we can test their consistency between infected host populations.

We apply our model to *gag* sequence data taken from three cohorts sampled from three continents: HOMER (British Columbia, Canada) [33], SSITT (Switzerland) [14, 34], and Bloemfontein (Bloemfontein, South Africa) [35, 36]. We find agreement between escape rate estimates, with the strongest positive correlation observed between CTL immune escape estimates across cohorts. Escape rate estimates are consistent between viral populations with distinct HIV-1 subtypes, where *gag* sequence divergence is as high as $\sim 9.2\%$. We also find consistent reversion rate estimates for the Bloemfontein and HOMER datasets (distinct HIV-1 subtypes, average nucleotide sequence divergence: 8.9%). Given the potential collection of paths through sequence space which could restore replicative fitness for a given escape variant, we find it particularly surprising to observe such a highly significant association between reversion rate estimates. These results suggest that given information regarding the dynamics of escape and reversion in one region, we may begin to make valid statements about these rate processes in other parts of the world, even when viral populations and HLA frequencies are substantially different. Our results support the findings in [37], in which escape pathways are found to be broadly similar even across viral subtypes, and align with fitness effects in Pol observed by Hinkley *et al.* [38] which demonstrate that whilst HIV-1 protease and reverse transcriptase is 'characterised by strong epistasis', a large portion of predictive power of replicative capacity is through single locus effects. To determine the extent of single locus effects on viral escape/reversion rates, we turn to data gathered from *in vitro* assessments of viral replicative capacity [39]. These estimates use site directed mutagenesis to induce escape variants in an otherwise conserved viral background, so do not reflect epistatic interactions in pathways to escape. We ask: do we observe correlations between our population derived escape or reversion rates, and these *in vitro* replicative capacity estimates? The emergence of a new consensus viral sequence within a given host is determined by the balance between immune escape

and replicative fitness cost within that host's immunological background [1, 40–43]. Naively ignoring all epistatic interactions, viral diversity, and assuming that all CTL escapes are equally beneficial to the virus, we would expect an escape mutation which incurs a small cost to viral replicative capacity to rapidly fix in the host's viral population. Similarly, an escape mutant which dramatically reduces replicative capacity would be expected to take far longer to reach intra-host consensus (we may make similar intuitive statements regarding reversion). Remarkably, despite these vast simplifying assumptions informing this intuition, we find a significant positive correlation between our SSITT escape rate estimates and *in vitro* replicative capacity, and a significant negative correlation between our HOMER reversion rate estimates and *in vitro* replicative capacity. We find it surprising to see such correlations between rate estimates measured in a reductionist, *in vitro* replicative capacity assay and a real world population of diverse individuals sharing diverse viruses. We conclude that whatever epistatic effects may be present across the HIV-1 proteome, in general rates of escape and reversion are largely dictated by the costs and benefits of individual mutations.

Methods

Cohorts

We consider paired viral sequence and host HLA type data taken from studies carried out in three populations: Switzerland (Swiss portion of the Swiss-Spanish intermittent treatment trial (SSITT) cohort [14, 34]); British Columbia, Canada (HAART observational medical evaluation and research (HOMER) cohort [33]); and Bloemfontein, South Africa [35, 36]. For each dataset bulk sequencing data was available, in which sequences obtained are assumed to represent the intra-host viral consensus sequence. We briefly outline the studies and portions of datasets used in Table 1. For each dataset, we restrict our attention to *gag* and to the subset of sequences with the majority subtype in each population as assessed using RIP [44]: subtype B for SSITT and HOMER, and subtype C for Bloemfontein. Our reasons are threefold. Firstly, non-synonymous mutations in *gag* are likely to have a detrimental fitness effect as they encode structural proteins that are among the most highly conserved in the viral proteome [45]. Indeed, variants in *gag* can result in a measurable reduction in viral replicative capacity, as estimated by a variety of methods [39, 46–49]. Secondly, the most protective HLA alleles are associated with CTL responses to Gag proteins [50–52]. Finally, site directed mutagenesis followed by replicative capacity assays have been carried out for a large number of known escape variants located in the *gag* gene [39].

Due to issues of confidentiality, we were unable to obtain the year of sequencing for the analysed portion of the HOMER dataset [33]. However, this is not required as we do not require a rescaling to calendar time (measured in *years*) from *substitutions site*$^{-1}$ to determine the existence of rank correlations of rate estimates across datasets. The 184 viral sequences which we analyse from HOMER dataset are taken from the year between 1996 and 1999 inclusive in which the largest number of *gag* sequences were obtained. For the SSITT dataset, separate collections of sequences were available for sequences encoding the p17 and p24 proteins in the *gag* gene. There was not a strong overlap between the patient identifiers of these two sequence sets, so we chose to analyse the two viral sequence regions independently.

Estimating escape and reversion rates

We estimate HLA prevalence using data from the HLA FactsBook [54]. For the Canadian (HOMER) and Swiss (SSITT) datasets, we set the prevalence at the 'Caucasian' estimate. For the African data (Bloemfontein, and the collection of cohorts summarised in Table S1), we set the prevalence at the 'Black' estimate. Epitopes examined by Boutwell *et al.* represent all optimally defined CTL epitopes at the time of publication [32]. Escape mutations within these epitopes were defined based upon a list of HLA-associated HIV polymorphisms identified via statistical association [22, 55]. We define escape

Dataset	SSITT [14, 34]	HOMER [53]	Bloemfontein [35, 36]
Population analysed	Switzerland	British Columbia, Canada	Bloemfontein, South Africa
Cohort size	$n = 133$	$n = 567$	$n = 884$,
	Zürich ($n = 29$), Geneva ($n = 26$),		plasma taken from $n = 278$
	Bern ($n = 11$), Basel ($n = 11$)		
	St Gall ($n = 8$), Lugano ($n = 7$)		
	Lausanne ($n = 5$)		
Sampling date	2000	Between 1996 and 1999	February - September 2006
Sequences analysed	p17 $n = 38$, p24 $n = 55$	$n = 184$	$n = 198$
Treatment	HAART	ART naive on recruitment, initiated HAART between August 1996 and September 1999	ART naive
Study requirements	Undetectable VL for >6 months, CD4$^+$ count $> 300 \mu l^{-1}$. no history of non-nucleoside reverse transcriptase inhibitors	≥ 3 antiretroviral drugs	Chronic, plasma taken by CD4$^+$ count, low and high favoured: 96 high ($> 500 \mu l^{-1}$), 18 medium ($200 - 400 \mu l^{-1}$), 164 low ($< 100 \mu l^{-1}$)

Table 1: Summary of three studies from which *gag* sequence data was available. For the SSITT and Bloemfontein datasets the number of sequences analysed is lower than the cohort size through a combination of lack of sequencing of *gag*, and data cleaning. For the HOMER dataset, the number of sequences analysed is lower than the cohort size due to restriction of the analysis to a single (unknown) calendar year, and data cleaning.

as any amino acid changes which occur at the same position as escape mutations provided in Boutwell *et al.* [39]. This is clearly an imperfect definition, but provides a sensible compromise between the clear overestimation of allowing all mutations within an epitope to be defined as escape, and only considering exact mutations validated as escapes *in vitro* [23]. In both of the approaches that we describe, it is important to note that estimates of escape represent averages across hosts known to restrict the epitope, and do not condition on the host making a CTL response or driving an escape during the course of infection. We are thus implicitly incorporating a scaling - the probability that an epitope is targetted conditional on the individual being HLA matched, which may vary across epitopes.

Maximum *a posteriori* (MAP) estimate using our integrated method

Full details of the method are provided in Palmer *et al.* [24]. Briefly, we combine Felsenstein's tree peeling algorithm with existing phylogenetic software to merge statistical phylogenetic techniques with an ordinary differential equation (ODE) model which captures the processes of transmission, escape, and reversion. By doing so, we are able to use the information contained within the viral genealogy to infer estimates of escape and reversion rates ($\lambda_{\mathrm{esc}}, \lambda_{\mathrm{rev}}$). There are four key steps in our inference regime:

1. Make the mild assumption that the genealogy and HLA/escape information are conditionally independent given the viral sequence information with the epitope removed.

2. We perform Markov chain Monte Carlo (MCMC) to sample genealogies from the posterior conditional on the viral sequence data with the epitope removed using BEAST [56]. We use a coalescent prior in an exponentially growing infected host population. The growth rate parameter of the infected population is sampled within the MCMC.

3. For each sampled viral genealogy, we determine the posterior density for ($\lambda_{\mathrm{esc}}, \lambda_{\mathrm{rev}}$). This is achieved through a modification of Felsenstein's peeling algorithm [57]. Further details of this step can be found in the Supplementary Methods.

4. By averaging over tree specific posteriors, we obtain a consensus posterior density for ($\lambda_{\mathrm{esc}}, \lambda_{\mathrm{rev}}$).

Purely ODE based, naive model

To analyse a particularly large fourth dataset for which rate estimation under our integrated model is not computationally feasible, we turn to a simpler compartment based ordinary differential equation (ODE) model of Fryer *et al.*. For each escape variant under consideration, viral sequence/HLA genotype pairs are split into four classes: HLA mismatched without escape, HLA mismatched with escape, HLA matched without escape and HLA matched with escape. Given these four proportions for an escape site, we then determine a best fit for λ_{esc} and λ_{rev} under the model described in Fryer *et al.* [23] during the exponential growth phase of the epidemic (Equations (1) - (4) in the Supplementary methods). Best estimates are determined using BFGS [58], an approximation to Newton's method of hill climbing. Time of the start of the epidemic is set at the time to the most recent common ancestor of a BEAST run on the *gag* sequence data for each dataset.

Estimates of time to escape/reversion could not be obtained when no individuals in the cohort possessed the restricting HLA. When an escape mutation residing in an epitope reaches consensus in the population, the variant peptide is known as a negatope [53, 59]. Negatopes of HIV-1 subtype B as reported in Boutwell *et al.* [39] are highlighted in Tables S2 - S4 by asterisks.

In vitro **viral replicative capacity**

In Boutwell *et al.* [39], the fitness cost of an escape variant is approximated by determining the *in vitro* replicative capacity of the escape mutant inserted into a subtype B lab strain, in the absence of any other sequence variation. The measure of fitness is essentially the exponential *in vitro* growth rate of an isolate relative to the lab strain. Full details of the procedure are provided in [39]. We use these published replicative capacity assay measurements for 23 *gag* escape variants where both viral sequence data and the relevant HLA locus information was available in all three cohorts.

Accounting for multiple testing

To account for multiple testing for each collection of tests of association, we perform permutation tests by shuffling the labellings of escape variants and re-evalaute correlation coefficients 100,000 times. This allows us to estimate the probability of observing a collection of Spearman rank correlation coefficients at least as extreme as those observed in the data. In all cases we obtain significant *p*-values. In each section, we also report the pairwise correlation coefficients, and describe a correlation as significant if its associated *p*-value is below 0.05.

Results

We apply our integrated model [24], and purely ODE based model [23] to data from the SSITT, HOMER and Bloemfontein cohorts. Detailed results, together with estimated replicative capacity of escape variants as previously estimated by an *in vitro* assay [39] are shown in Tables S2, S3 and S4 for the HOMER, SSITT and Bloemfontein datasets respectively. Figure S1 summarises the prevalence of escape in all three regions. For each escape site we determine the proportion of HLA matched hosts harbouring the viral escape form. We perform the same calculation for HLA mismatched hosts. These two values for each of the 23 escape sites are then plotted against each other. The majority of the points lie below $y = x$, suggesting that selection for escape is taking place in the majority of the HLA/epitope pairings.

Estimates of time to escape and time to reversion for the HOMER dataset are given in units of *substitutions site*$^{-1}$. This is because the date of sampling was not available due to issues of confidentiality, and thus a scaling from *substitutions site*$^{-1}$ to *years* could not be estimated using BEAST [60]. Throughout our results, we report the Spearman rank correlation coefficient. We choose this statistic as variance
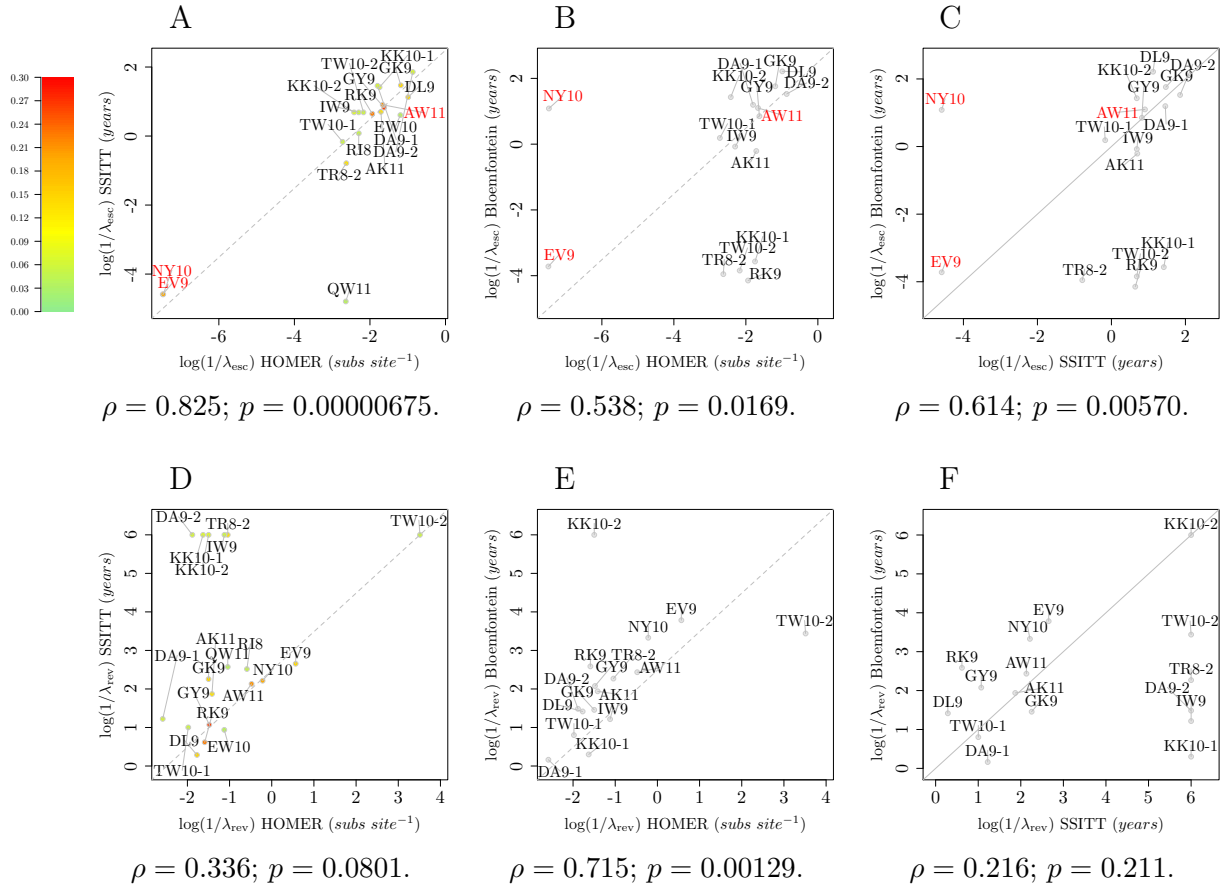
Figure 1: **Comparing estimated times to escape/reversion in each pair of cohorts**. Figures 1A - 1C show MAP estimates of time to escape for the SSITT vs. HOMER, Bloemfontein vs. SSITT and Bloemfontein vs. HOMER MAP estimates respectively. Figures 1D - 1F show the corresponding MAP estimates of times to reversion. Numbers following a dash refer to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Negatopes are highlighted in red for time to escape. The scale bar indicates the HLA prevalence taken from the HLA handbook [54] for each of the restricting HLA types associated to escape in Caucasian populations. $y = x$ is plotted in solid grey where estimates are on the same timescale. Dotted grey lines represent an estimate of $y = x$ after a change of timescale assuming the Swiss and Canadian epidemics are expanding at roughly the same rate.

in rate estimates increases drastically as the true underlying rates tend to infinity (or zero if we use a log scale). Also, when comparing replicative capacity to rate of escape or reversion, it is unclear *a priori* that any such relationship would be linear. In each plot, epitopes are abbreviated by writing the first amino acid of the epitope, followed by last, followed by the epitope length (e.g KRWIILGLNK → KK10).

Are rates of escape and reversion consistent across infected populations in different parts of the world?

Escape rates

Figure 1A-1C shows scatter-plots of estimates of average time to escape for the different combinations of populations: HOMER vs. SSITT, HOMER vs. Bloemfontein, and SSITT vs. Bloemfontein respectively. We find strong positive correlations between escape rate estimates for the SSITT and HOMER cohorts ($\rho = 0.825$; $p = 0.00000675$). We do not find clustering of restricting HLAs of similar prevalence in any portion of the space of times to escape suggesting that the observed correlation is

not an artefact of the HLA prevalence in the two populations. We also observe significant positive correlations between estimates of average time to escape for the remaining population combinations in which HIV-1 subtypes are distinct (analysed SSITT and HOMER viral sequences are subtype B, analysed Bloemfontein viral sequences are subtype C): HOMER vs. Bloemfontein: $\rho = 0.538$; $p = 0.0169$, and SSITT vs. Bloemfontein: $\rho = 0.614$; $p = 0.00570$. These results suggest that escape rates are comparable across infected host populations. Under a permutation test (see Methods) we find that the results are significant, $p < 0.00001$.

In determining the strength of positive correlation between escape rate estimates across populations, we note that a potential source of bias is the inclusion of negatopes: any mutation at high prevalence across hosts will result in a high escape rate estimate, and consequently may lead to spurious associations. To determine if this is the case in our data we remove negatopes (highlighted in red in Figure 1), and re-evaluate correlation coefficients and p-values. All positive correlations remained significant after removal of negatopes (HOMER vs. SSITT: $\rho = 0.779$; $p = 0.000189$, HOMER vs. Bloemfontein: $\rho = 0.549$; $p = 0.0276$, and SSITT vs. Bloemfontein: $\rho = 0.657$; $p = 0.00730$). The p-value for the permutation test to account for multiple testing also remained significant, $p = 0.0006$.

We sought to validate our findings with further data. A potential source of error in our Bloemfontein escape rate estimates was the low number of individuals in the Bloemfontein dataset with the restricting HLA (shown in Table S3), due to the rarity of these HLA genotypes in African populations coupled with the limited size of the dataset. Additionally, some escape mutations are found at 100% prevalence or are not present at all in the Bloemfontein dataset. 9/20 rate estimates are obtained from data for which just one individual has the restricting HLA, or 0%/100% of HLA matched hosts have the escape mutation. Such escape proportions in HLA matched hosts result in a lack of power to estimate escape rates and can lead to spurious rate estimates [24] (they may lead to estimates of $\sim$ instantaneous escape or $\sim$ infinite time to escape). To attempt to correct for this source of error we obtained a far larger dataset of African sequences coupled with host HLA genotype data [35, 61–67]. Unfortunately, such a large dataset was computationally intractable using our integrated model. We therefore turn to the simpler ODE model of Fryer *et al.*, noting the agreement between escape rate estimates derived with our integrated method and this compartment based ODE approach [24], see Table S5 and Palmer *et al* [24]. The large dataset summarised in Table S1 is not a single study in a single homogeneous mixing population. Nevertheless, we apply the ODE method to all of the data in Table S1. Scatter-plots of estimates of time to escape using the naive ODE method on the large African dataset against the two HIV-1 subtype B datasets are shown in Figures S3. In both cases we find significant positive correlations ($\rho = 0.711, p = 0.000963$ and $\rho = 0.765$; $p = 0.000175$ for African against HOMER and African against SSITT time to escape estimates respectively) which remained significant after removal of negatopes (African vs. HOMER $\rho = 0.723$; $p = 0.00240$, African vs. SSITT $\rho = 0.819$; $p = 0.000170$). We also note that the strongest disagreement seen in the plots occurs when time to escape is very low in one or other of the populations. This occurs when the escape mutation is at high prevalence in one or other of the populations, possibly as a result of founder effects.

Reversion rates

Figures 1D-1F and S4 show the corresponding scatter-plots of estimates of time to reversion. We find a significant agreement between the rank ordering of HOMER and Bloemfontein reversion rate estimates ($\rho = 0.715, p = 0.00129$). This correlation was not repeated in any of the other combination of the analysed cohorts. However, the collection of correlation coefficients observed are significant under a permutation test, $p = 0.0011$. We note that TLYCVHQR, [A]ISPRTLNAW, DRFYKTLRA, KRWIILGLNK, and KRWIILGLNK are estimated to have extremely low reversion rates when using the SSITT data. This is potentially due to the small SSITT cohort size and the inherent difficulty in estimating reversion rates when the restricting HLA type is rare and/or the true underlying reversion rate is low (see simulation studies in [24]). Further data is required to validate these rate estimates.

Sequence divergence within and between cohorts

The consistency between escape rate estimates in three different populations seems to suggest that the viral background on which an escape, or pathway to escape is placed is generally a secondary effect during CTL immune evasion. If it weren't, then we would be expect such effects to mask positive correlations between escape rate estimates across host populations. To investigate the extent of viral divergence between the examined datasets, we determined nucleotide and amino acid sequence divergence as measured by Hamming distance (Figure S7). We evaluate this metric for the region of *gag* over which we had the largest number of available sequences. This was to avoid biasing through differences in sequence diversity across *gag*. A neighbour joining tree using the K81 model [68] is shown in Figure S8. Taken together, Figures S7 and S8 show a dramatic divergence between the subtype B and C sequences, as we would expect. We observe mixing of lineages between Canada and Switzerland, coupled with examples of cohort specific clades.

The consistent and significant positive association between escape rate estimates exists despite a mean divergence of up to 9.2% between viral sequence datasets.

Escape and reversion rates within cohorts vs. *in vitro* replicative capacity estimates

To examine the importance of single escape variants upon rates of escape and reversion, we consider the most extreme case: complete removal of all sequence variation except at the site of the escape variant. We compare *in vitro* replicative capacity of site directed mutants with our escape and reversion estimates. For a collection of 23 epitope variants we had access to sequence data and an estimate of replicative fitness cost from Boutwell *et al.*, in which site directed mutagenesis was applied to a subtype B lab strain. If we naively assume an absence of epistasis and that the benefit of evading CTL mediated immunity is equal for all escape variants, then we would expect to see average time to escape negatively correlated with *in vitro* replicative capacity, and average time to reversion positively correlated with *in vitro* replicative capacity.

Subtype B: Escape and reversion rate estimates in SSITT and HOMER datasets vs. *in vitro* replicative capacity estimates

We first examine the correlation between this *in vitro* measure of replicative capacity and rate estimates in the two populations in which subtype B is the predominant subtype. Results are displayed in Figure 2 (permutation test p-value to account for multiple testing of escape rate versus replicative fitness; $p = 0.00192$, corresponding p-value for reversion; $p = 0.0113$). Surprisingly, despite these naive assumptions we indeed observe a negative correlation between time to escape and replicative capacity for the Spearman rank correlation coefficients which reaches significance for the SSITT data ($\rho = -0.451$, $p = 0.0231$). We also observe a positive correlation between time to reversion and replicative capacity which reaches significance for the HOMER data ($\rho = 0.530$, $p = 0.00815$). These findings again suggest that epistatic effects are generally secondary to either the benefits to the virus of CTL escape in HLA matched hosts, or the detriment to viral replicative capacity in HLA mismatched hosts. Such effects appear to be overwhelmed by the impact of insertion or removal of escape mutants. Logically, if epistasis dramatically affected the available pathways to escape/reversion then we would not expect to see time to escape or reversion correlate (negatively and positively respectively) with viral replicative capacity experiments which discard all sequence variation aside from the escape mutant. Finally, we note that outliers in Figure 2C are exactly those discussed in the previous section. Again we emphasise that larger European sequence datasets with host HLA information should be analysed to determine the accuracy of the reversion rate estimates for these outliers.
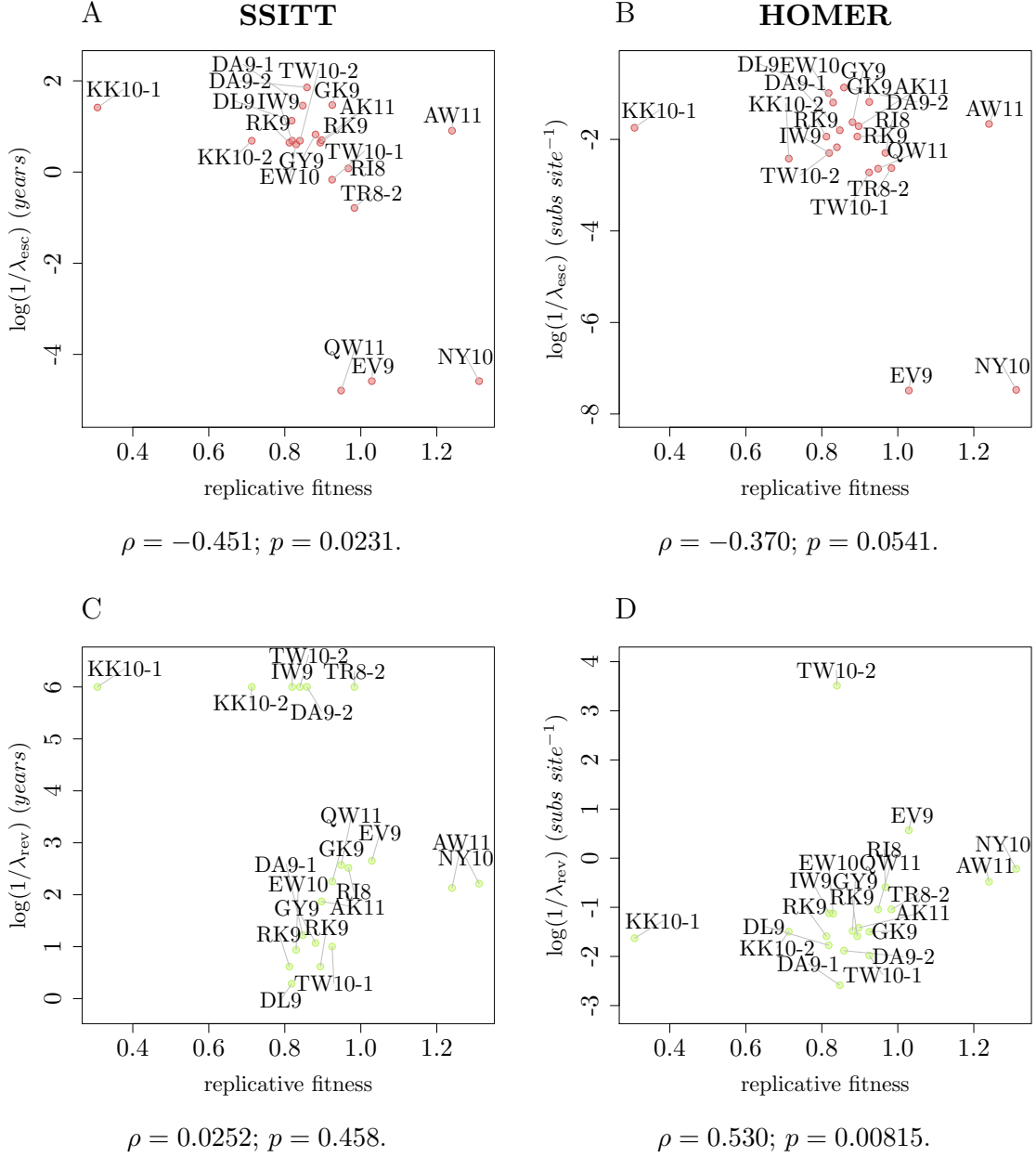
Figure 2: **Replicative capacity plotted against the MAP estimates of time to escape and reversion**. The first and second columns display estimates obtained using *in vivo* data from Switzerland (SSITT) and Canada (HOMER) respectively. Numbers after the epitope abbreviation refer to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Time to escape is measured in *years* except in the case of the HOMER cohort in which it is measured in units of substitutions. Spearman rank correlation coefficients; ρ, with associated *p*-values are displayed on each plot.
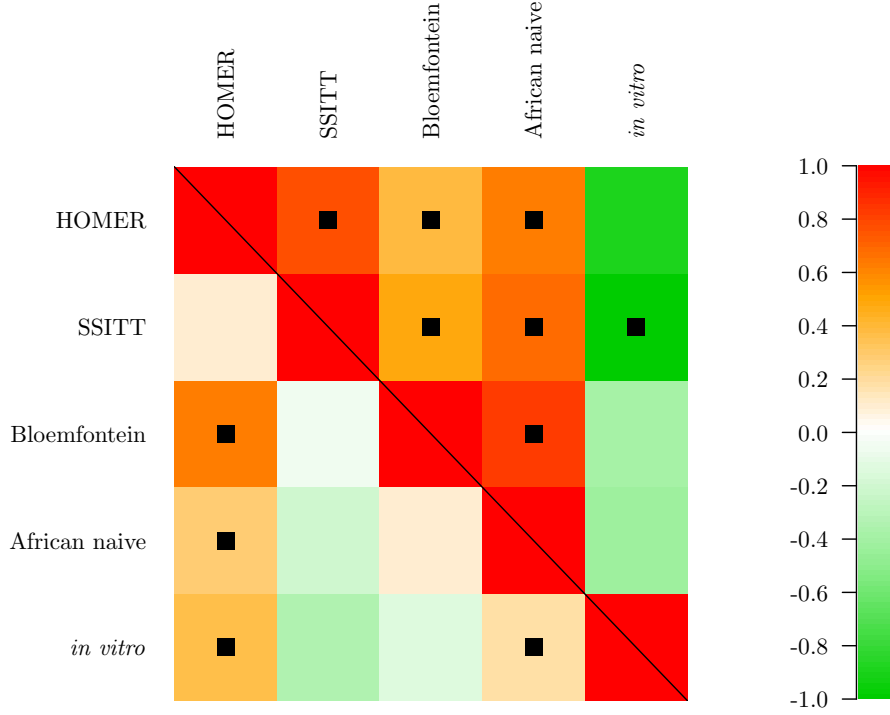
Figure 3: **Summary of all associations**. Spearman rank correlations between expected time to escape across cohorts and with *in vitro* replicative capacity are shown above the leading diagonal. Spearman rank correlations between reversion rates across cohorts and with *in vitro* replicative capacity are shown below the leading diagonal. Spearman rank correlation coefficients with $p < 0.05$ are highlighted with a black square. Associated Spearman rank correlation coefficients are coloured according to the colourbar.

Subtype C: Escape and reversion rate estimates in the Bloemfontein dataset vs. *in vitro* **replicative capacity estimates**

We next sought to determine if this correlation was replicated in the Bloemfontein dataset where subtype C predominates. We do not find any such correlation between estimated time to escape/reversion in Bloemfontein (or the full African dataset using the naive ODE method, see Figure S6) and *in vitro* replicative capacity. We note that the permutation test p-value remains significant after accounting for these further comparisons, $p = 0.0113$.

Discussion

There is a large literature detailing the existence and complexity of pathways to escape, compensatory mutations and other forms of epistasis in HIV-1 [4, 21, 22, 38, 43, 69–73]. Given this complexity, it is unclear whether the processes of CTL immune escape and reversion can be meaningfully represented by describing each escape variants rate of escape and reversion. In this work, we sought to determine if it makes sense to characterise an escape variants with an escape and reversion rate, and ask if such rate estimates are consistent across host populations. We found significant consistency between expected time to escape for each cohort combination, even across HIV-1 subtypes where average sequence divergence between viral populations was as high as $\sim 9.2\%$. Consistency in escape/reversion

rate estimates across populations lead to the following question: is the fitness effect at a single locus a predictor of the tempo of viral escape or reversion? To test this, we considered estimates of *in vitro* replicative capacity applied to a subtype B lab strain [39]. If we naively assume that all escape mutations have the same impact upon CTL immune evasion, then we would expect CTL escape mutations that are more costly *in vitro* (i.e. that have a lower viral replicative capacity) to escape more slowly in HLA matched hosts and revert more rapidly in HLA mismatched hosts. The use of *in vitro* replicative capacity of single escape variants applied to a lab strain as a measure of its fitness averaged over an entire infected population is clearly a vast simplification. Our method also requires a number of simplifying assumptions [24]. Despite this, as expected, we find a significant negative correlation between *in vitro* replicative capacity and estimated time to escape in the Swiss dataset (SSITT), and a significant positive correlation *in vitro* replicative capacity and *in vivo* estimates of time until reversion in the HOMER dataset. For the Bloemfontein dataset (consisting of subtype C HIV-1 sequences), no association was found for either escape or reversion estimates. This could suggest a role for epistatic interactions, potentially involving sites distinguishing subtypes B and C. However, we note that the similar replicative capacity of measured escape variants and the noise of such assays, coupled with our low power to estimate many of these rates given the rarity of the associated HLA types in African populations could also lead to a loss of signal of association. Furthermore, founder effects and overlapping epitopes associated to other HLA types have the potential to affect our rate estimates.

Shannon entropy [74] obtained from viral sequences taken across infected hosts has been used as a surrogate measure of the replicative fitness cost of CTL escape [31] as it provides a metric for the variability at a given site. Authors thus assume that a site of lower variability will escape more slowly. However, whilst useful, this simple metric ignores the relative prevalence of HLA types in the host population, the dependency structure inherent in the viral genealogy, and has no notion of 'escape' in its calculation. By using a simple model of escape and reversion across infected hosts to pick apart the relative contributions to variation observed at escape sites we obtain biologically meaningful parameters which correlate with replicative fitness as we would (perhaps naively) expect. For all of our analysed datasets, Shannon entropy does not correlate with viral replicative capacity ($\rho = 0.193$; $p = 0.208$, $\rho = 0.0226$; $p = 0.462$, $\rho = 0.0850$; $p = 0.361$, for the SSITT, HOMER, and Bloemfontein cohorts respectively).

During the process of escape, it appears that epistatic effects are often overpowered by the effect of a single point mutation; giving rise to consistent escape rate estimates across host populations. This result is powerful as it allows us to inform a null hypothesis when studying a new HIV-1 dataset. Evidence for correlation between reversion rate estimates across populations and with *in vitro* replicative capacity is less clear cut. This may be due to the inherent difficulty in estimating reversion rates. Indeed, there is contradiction in the literature regarding the rate of reversion. A few famous epitopes consistently revert in the absence of HLA pressure [15,20,31,62]. However, whether this readily extends across escape variants is a subject of debate. Some cohort studies have suggested little evidence for reversion [5, 31, 75], and cross-sectional estimates suggest a low rate of reversion on average [23, 24]. However, a large longitudinal study [76], transmission pair data [77] and separate cohort studies [45], have suggested that reversion occurs early and is widespread. Other studies have concluded that between 20% and 30% of non-synonymous substitutions within host constitute reversions [45,78,79]. Such estimates depend on the authors' definition of reversion, and highlight the need for data sharing (both host HLA and viral sequence data) to critically compare methodologies, increase statistical power to inform estimates and increase our understanding of the underlying biology.

Under our cross-sectional model, reliable reversion rate estimation requires multiple instances of transmission of escape mutations followed by reversion in an HLA mismatched host, an event which occurs less often than transmission of wildtype at the site(s) under consideration (particularly if the restricting HLA is rare, or the escape rate is relatively low). Indeed, simulation studies show that a lack of data and low underlying reversion rates can lead to dramatic underestimation of the true rate [24]. Despite this, we observe a strong signal of association between the reversion rates estimated using the Bloemfontein data and those estimated using the HOMER dataset. An alternative explanation for the

lack of a signal of correlation in reversion rate estimates between cohorts is that epistatic effects are important in a subset of the analysed epitopes. Epistatic effects playing a stronger role in the reversion process does make sense intuitively: on average a viral lineage spends a longer period of time within HLA mismatched hosts (because HLA diversity is so high). Selection to increase replicative capacity acts across the entirety of the viral genome both in the presence and absence of any HLA associated immune pressure (contrast this to CTL associated selective pressure, which is strongest in and around epitopes). The high dimensionality of sequence space would then lead us to expect multiple pathways to a restoration of viral fitness, not all of which may be associated to a simple reversion of the purported escape mutation - particularly if the 'survival of the flattest' hypothesis is borne out in reality [80].

This study has a number of limitations. Throughout we have applied our methods to subtype B and C viral sequence data from Canada, Switzerland, South Africa, Malawi and Botswana. Access to further host HLA information coupled with cross-sectional HIV-1 sequence data would allow us to determine if similar orderings of escape and reversion rates are observed in further infected host populations, potentially harboring viruses with other HIV-1 subtypes. This will be particularly interesting given the dissimilar escape pathways found by Avila-Rios *et al* when analysing a collection of subtype B sequences sampled from Mexico [71], and the significant number of further unique HLA associated amino acid variants in other viral subtypes and circulating recombinant forms across the globe [81]. The escape variants analysed here are all found in *gag*. It would therefore be interesting to determine whether our results may be extended to other portions of the HIV-1 genome. Reliably interpreting the effect of replicative capacity on escape and reversion rates in less conserved regions will be difficult. Although replicative fitness assays are improving all the time [39, 46–48], obtaining the required sensitivity to distinguish between small differences in replicative capacity associated to variants in less conserved regions of the viral genome is no easy task. For example, small replicative fitness costs known reverting mutations can often not be distinguished [43]. The majority of the collection of optimal epitopes studied [32] have been determined using data from European and North American populations, and thus associated HLA types are likely to be found at high or moderate frequencies within these populations. However, differences in HLA allele frequency distributions across the world often makes rate estimation using data taken from other parts of the world difficult - we sometimes require far larger datasets to obtain rate estimates for escape variants in one region versus another. Datasets with $> \sim 500$ sequences in HIV-1 *gag* become intractable using our integrated method. Ability to incorporate vast host/virus datasets would allow us to determine escape and reversion rate estimates for escape variants associated to rare HLA alleles and obtain more accurate estimates for escape mutations associated to more common alleles whilst accounting for dependency structure inherent in viral sequence data.

Our results are summarised in Figure 3. Escape rates are consistent across three populations in three different continents and two viral subtypes. We also demonstrate that reversion rates correlate between estimates obtained from a Canadian and African datasets, though further data are required to determine if this observation is true in general.

Acknowledgements

References

[1] Goulder PJR, Watkins DI (2004) HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol* 4(8):630–640.

[2] McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF (2010) The immune response during acute HIV-1 infection: clues for vaccine development. *Nature reviews. Immunology* 10(1):11–23.

[3] Roberts HE et al. (2015) Structured Observations Reveal Slow HIV-1 CTL Escape. *PLoS Genet* 11(2):e1004914+.

[4] Troyer RM et al. (2009) Variable Fitness Impact of HIV-1 Escape Mutations to Cytotoxic T Lymphocyte (CTL) Response. *PLoS Pathog* 5(4):e1000365+.

[5] Sunshine JE et al. (2015) Fitness-balanced escape determines resolution of dynamic founder virus escape processes in HIV-1 infection. *Journal of Virology* pp. JVI.01876–15+.

[6] Alberts B et al. (2007) *Molecular Biology of the Cell: Reference Edition.* (Garland Science), 5 edition.

[7] Robinson J et al. (2013) The IMGT/HLA database. *Nucleic Acids Research* 41(D1):D1222–D1227.

[8] Zhang SCC et al. (2012) Aminopeptidase substrate preference affects HIV epitope presentation and predicts immune escape patterns in HIV-infected individuals. *Journal of immunology (Baltimore, Md. : 1950)* 188(12):5924–5934.

[9] Draenert R et al. (2004) Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *The Journal of experimental medicine* 199(7):905–915.

[10] Iversen AK et al. (2006) Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nature immunology* 7(2):179–189.

[11] Goulder PJR, Walker BD (2012) HIV and HLA Class I: An Evolving Relationship. *Immunity* 37(3):426–440.

[12] Allen TM et al. (2004) Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *Journal of virology* 78(13):7069–7078.

[13] Frater AJ et al. (2006) Passive Sexual Transmission of Human Immunodeficiency Virus Type 1 Variants and Adaptation in New Hosts. *J. Virol.* 80(14):7226–7234.

[14] Frater AJ et al. (2007) Effective T-Cell Responses Select Human Immunodeficiency Virus Mutants and Slow Disease Progression. *Journal of Virology* 81(12):6742–6751.

[15] Leslie AJ et al. (2004) HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med* 10(3):282–289.

[16] Martinez-Picado J et al. (2006) Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1. *Journal of Virology* 80(7):3617–3623.

[17] Davenport MP, Loh L, Petravic J, Kent SJ (2008) Rates of HIV immune escape and reversion: implications for vaccination. *Trends in microbiology* 16(12):561–566.

[18] Schneidewind A et al. (2007) Escape from the Dominant HLA-B27-Restricted Cytotoxic T-Lymphocyte Response in Gag Is Associated with a Dramatic Reduction in Human Immunodeficiency Virus Type 1 Replication. *Journal of Virology* 81(22):12382–12393.

[19] Brockman MA et al. (2007) Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A. *Journal of Virology* 81(22):12608–12618.

[20] Crawford H et al. (2007) Compensatory Mutation Partially Restores Fitness and Delays Reversion of Escape Mutation within the Immunodominant HLA-B*5703-Restricted Gag Epitope in Chronic Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology* 81(15):8346–8351.

[21] Kelleher AD et al. (2001) Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *The Journal of experimental medicine* 193(3):375–386.

[22] Brumme ZL et al. (2009) HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins. *PLoS ONE* 4(8):e6687+.

[23] Fryer HR et al. (2010) Modelling the Evolution and Spread of HIV Immune Escape Mutants. *PLoS Pathog* 6(11):e1001196+.

[24] Palmer D, Frater J, Phillips R, McLean AR, McVean G (2013) Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes. *Proceedings of the Royal Society B: Biological Sciences* 280(1762):20130696+.

[25] Asquith B, Edwards CTT, Lipsitch M, McLean AR (2006) Inefficient Cytotoxic T Lympho-cyte–Mediated Killing of HIV-1–Infected Cells In Vivo. *PLoS Biol* 4(4):e90+.

[26] Asquith B, McLean AR (2007) In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proceedings of the National Academy of Sciences* 104(15):6365–6370.

[27] Fernandez CS et al. (2005) Rapid Viral Escape at an Immunodominant Simian-Human Immun-odeficiency Virus Cytotoxic T-Lymphocyte Epitope Exacts a Dramatic Fitness Cost. *Journal of Virology* 79(9):5721–5731.

[28] Ganusov VV, De Boer RJ (2006) Estimating Costs and Benefits of CTL Escape Mutations in SIV/HIV Infection. *PLoS Comput Biol* 2(3):e24+.

[29] Goonetilleke N et al. (2009) The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *The Journal of Experimental Medicine* 206(6):1253–1272.

[30] Ganusov VV et al. (2011) Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV infection. *Journal of virology* 85(20):10518–10528.

[31] Liu MK et al. (2013) Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of clinical investigation* 123(1):380–393.

[32] Llano A, Williams A, Olvera A, Silva-Arrieta S, Brander C (year?) Best-characterized hiv-1 ctl epitopes: The 2013 update.

[33] Brumme ZL et al. (2008) Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS (London, England)* 22(11):1277–1286.

[34] Fagard C et al. (2003) A prospective trial of structured treatment interruptions in human immun-odeficiency virus infection. *Archives of internal medicine* 163(10):1220–1226.

[35] Huang KHGH et al. (2009) Prevalence of HIV type-1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. *Antiviral therapy* 14(7):975–984.

[36] Huang KHG et al. (2011) Progression to AIDS in South Africa Is Associated with both Reverting and Compensatory Viral Mutations. *PLoS ONE* 6(4):e19018+.

[37] Carlson JM et al. (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS computational biology* 4(11):e1000225+.

[38] Hinkley T et al. (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* 43(5):487–489.

[39] Boutwell CL et al. (2013) Frequent and Variable Cytotoxic-T-Lymphocyte Escape-associated Fitness Costs in the Human Immunodeficiency Virus Type 1 Subtype B Gag Proteins. *Journal of Virology* 87(7):3952–3965.

[40] Liu Y et al. (2007) Evolution of human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitopes: fitness-balanced escape. *Journal of virology* 81(22):12179–12188.

[41] Troyer RM et al. (2009) Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS pathogens* 5(4).

[42] Allen TM, Altfeld M (2008) Crippling HIV one mutation at a time. *The Journal of Experimental Medicine* 205(5):1003–1007.

[43] Song H et al. (2014) Reversion and T Cell Escape Mutations Compensate the Fitness Loss of a CD8+ T Cell Escape Mutant in Their Cognate Transmitted/Founder Virus. *PLoS ONE* 9(7):e102734+.

[44] Siepel AC, Halpern AL, Macken C, Korber BT (1995) A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS research and human retroviruses* 11(11):1413–1416.

[45] Li B et al. (2007) Rapid Reversion of Sequence Polymorphisms Dominates Early Human Immunodeficiency Virus Type 1 Evolution. *Journal of Virology* 81(1):193–201.

[46] Lanxon-Cookson EC et al. (2013) Factors affecting relative fitness measurements in pairwise competition assays of human immunodeficiency viruses. *Journal of Virological Methods* 194(1-2):7–13.

[47] Song H et al. (2012) Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. *Retrovirology* 9(1):89+.

[48] Liu Y et al. (2013) A sensitive real-time PCR based assay to estimate the impact of amino acid substitutions on the competitive replication fitness of human immunodeficiency virus type 1 in cell culture. *Journal of virological methods* 189(1):157–166.

[49] Prince JL et al. (2012) Role of Transmitted Gag CTL Polymorphisms in Defining Replicative Capacity and Early HIV-1 Pathogenesis. *PLoS Pathog* 8(11):e1003041+.

[50] Kiepiela P et al. (2007) CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature medicine* 13(1):46–53.

[51] Zuñiga R et al. (2006) Relative Dominance of Gag p24-Specific Cytotoxic T Lymphocytes Is Associated with Human Immunodeficiency Virus Control. *Journal of Virology* 80(6):3122–3125.

[52] Novitsky V et al. (2003) Association between Virus-Specific T-Cell Responses and Plasma Viral Load in Human Immunodeficiency Virus Type 1 Subtype C Infection. *Journal of Virology* 77(2):882–890.

[53] Leslie A et al. (2005) Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *The Journal of experimental medicine* 201(6):891–902.

[54] Marsh SGE, Parham P, Barber LD (2000) *The HLA FactsBook.* (Academic Press), 1 edition.

[55] Wang YE et al. (2009) Protective HLA Class I Alleles That Restrict Acute-Phase CD8+ T-Cell Responses Are Associated with Viral Escape Mutations Located in Highly Conserved Regions of Human Immunodeficiency Virus Type 1. *Journal of Virology* 83(4):1845–1855.

[56] Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8):1969–1973.

[57] Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17(6):368–376.

[58] Nash JC (1990) *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* (CRC Press), 2 edition.

[59] Moore CB et al. (2002) Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science (New York, N.Y.)* 296(5572):1439–1443.

[60] Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7(1):214+.

[61] Kiepiela P et al. (2004) Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 432(7018):769–775.

[62] Kawashima Y et al. (2009) Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458(7238):641–645.

[63] Rousseau CM et al. (2008) HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *Journal of virology* 82(13):6434–6446.

[64] Leslie A et al. (2010) Additive Contribution of HLA Class I Alleles in the Immune Control of HIV-1 Infection. *Journal of Virology* 84(19):9879–9888.

[65] Matthews PC et al. (2011) HLA-A*7401–Mediated Control of HIV Viremia Is Independent of Its Linkage Disequilibrium with HLA-B*5703. *The Journal of Immunology* 186(10):1003711–5686.

[66] Shapiro RL et al. (2010) Antiretroviral Regimens in Pregnancy and Breast-Feeding in Botswana. *N Engl J Med* 362(24):2282–2294.

[67] Matthews PC et al. (2008) Central Role of Reverting Mutations in HLA Associations with Human Immunodeficiency Virus Set Point. *Journal of Virology* 82(17):8548–8559.

[68] Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America* 78(1):454–458.

[69] Carlson JM et al. (2012) Widespread Impact of HLA Restriction on Immune Control and Escape Pathways of HIV-1. *Journal of Virology* 86(9):5230–5243.

[70] Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in HIV-1. *Science (New York, N.Y.)* 306(5701):1547–1550.

[71] Avila-Rios S et al. (2009) Unique features of HLA-mediated HIV evolution in a Mexican cohort: a comparative study. *Retrovirology* 6(1):72+.

[72] McMichael AJ (2007) Triple bypass: complicated paths to HIV escape. *The Journal of experimental medicine* 204(12):2785–2788.

[73] Dahirel V et al. (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences* 108(28):11530–11535.

[74] Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal, The* 27(3):379–423.

[75] Gounder K et al. (2015) High Frequency of Transmitted HIV-1 Gag HLA Class I-Driven Immune Escape Variants but Minimal Immune Selection over the First Year of Clade C Infection. *PLoS ONE* 10(3):e0119886+.

[76] Duda A et al. (2009) HLA-Associated Clinical Progression Correlates with Epitope Reversion Rates in Early Human Immunodeficiency Virus Infection. *Journal of Virology* 83(3):1228–1239.

[77] Carlson JM et al. (2014) HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science (New York, N.Y.)* 345(6193):1254031+.

[78] Zanini F et al. (2015) Population genomics of intrapatient HIV-1 evolution.

[79] Allen TM et al. (2005) Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *Journal of virology* 79(21):13239–13249.

[80] Rolland M et al. (2007) HIV-1 over time: fitness loss or robustness gain? *Nat Rev Micro* 5(9):1–2.

[81] Gesprasert G et al. (2010) HLA-associated immune pressure on Gag protein in CRF01_AE-infected individuals and its association with plasma viral load. *PloS one* 5(6).

[82] Cromer D, Wolinsky SM, McLean AR (2010) How fast could HIV change gene frequencies in the human population? *Proceedings of the Royal Society B: Biological Sciences* 277(1690):1981–1989.

[83] Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* 14(7):717–724.

Supplementary Materials

Supplementary methods: Elaborating on step 3 of our inference regime.

After sampling a genealogy from the posterior output by BEAST, we use the host HLA and escape information at the tips of the genealogy to determine estimates of $(\lambda_{esc}, \lambda_{rev})$. Our algorithm is based on Felsenstein's tree pruning algorithm [57], where the tip information is of the form $(a, b); a, b \in \{0, 1\}$ such that $\{0, 1\}$ in the first entry denotes {HLA mismatch, HLA match}, and $\{0, 1\}$ in the second entry denotes {no escape mutation, escape mutation} at the epitope under investigation. Transmission is assumed to take place at rate λ. The probability that a given individual is HLA matched at the epitope under investigation is q (assumed to be constant in the population over time, as any selection on hosts would occur on far larger timescales [82]). We then determine the probability of a given configuration of the data at the tips, conditional on the genealogy, a collection of epidemiological parameters and a collection of parameters $\boldsymbol{\lambda} = \{\lambda_{esc}, \lambda_{rev}, q\}$, which model escape and reversion down each lineage governed by an instantaneous rate matrix $\mathbf{Q}$. The instantaneous rate matrix is based on the following system of ODEs:

HLA mismatched hosts infected with viruses without escape:

$$\frac{dY_0^0(t)}{dt} = \lambda(1 - q)\left(Y_0^0(t) + Y_0^1(t)\right) + \lambda_{rev}Y_1^0(t) - \mu Y_0^0(t) \tag{1}$$

HLA mismatched hosts infected with viruses with escape:

$$\frac{dY_1^0(t)}{dt} = \lambda(1 - q)\left(Y_1^0(t) + Y_1^1(t)\right) - \lambda_{rev}Y_1^0(t) - \mu Y_1^0(t) \tag{2}$$

HLA matched hosts infected with viruses without escape:

$$\frac{dY_0^1(t)}{dt} = \lambda q\left(Y_0^0(t) + Y_0^1(t)\right) - \lambda_{esc}Y_0^1(t) - \mu Y_0^1(t) \tag{3}$$

HLA matched hosts infected with viruses with escape:

$$\frac{dY_1^1(t)}{dt} = \beta c q\left(Y_1^0(t) + Y_1^1(t)\right) + \lambda_{esc}Y_0^1(t) - \mu Y_1^1(t) \tag{4}$$

where λ is the transmission rate. We assume that escape may only occur from the $(1, 0)$ state at rate λ_{esc} and reversion may only occur from the $(0, 1)$ state at rate λ_{rev}, as in As in [23, 24]. The

Dataset	Durban [64,67]	Mma Bana [66]	Kimberley [65]	Thames valley [65]
Population analysed	Durban, South Africa	Gabarone, Botswana	Kimberley, South Africa	Zimbabwe, South Africa, Malawi and Botswana.
Cohort size	$n = 1218$	$n = 514$	$n = 31$	$n = 102$
Sampling date	1999 - 2006	July 2006 - May 2008	–	–
Sequences analysed	$n = 929$	$n = 321$	$n = 26$	$n = 65$ Zimbabwe ($n = 55$), Malawi ($n = 7$) South Africa ($n = 3$).
Treatment	ART naive	ART naive	ART naive	ART naive
Study requirements	Recruited following voluntary counselling and testing in anti-natal or outpatient clinics.	Pregnant women	Postnatal mothers	Southern African subjects attending outpatient HIV clinics in the Thames Valley area of the UK

Table S1: Summary of four African studies from which *gag* sequence data and HLA information was available. In each case, the number of sequences analysed is lower than the cohort size through a combination of lack of sequencing of *gag*, and data cleaning. These data were combined and analysed using the ODE model.

instantaneous rate matrix $\mathbf{Q}$ may then be defined as

$$
\mathbf{Q} = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \\ \left(\begin{array}{cccc} -\hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q & 0 \\ \lambda_{\text{rev}} & -\lambda_{\text{rev}} - \hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q \\ \hat{\lambda}(t)(1-q) & 0 & -\lambda_{\text{esc}} - \hat{\lambda}(t)(1-q) & \lambda_{\text{esc}} \\ 0 & \hat{\lambda}(t)(1-q) & 0 & -\hat{\lambda}(t)(1-q) \end{array} \right), \end{array} \tag{5}
$$

where $\hat{\lambda}(t) = \lambda p_0(T_{\text{MRCA}} - t)$, with time increasing towards the present. T_{MRCA} is the time before the present of the most recent common ancestor (MRCA), when $t = 0$. Q_{ij} refers to the transition from state i to state j. $p_0(t)$ is the probability that a lineage at time t in the past does not have any sampled descendants.

$$
p_0(t) = 1 - \frac{\widetilde{\rho}(\lambda - \mu)}{\widetilde{\rho}\lambda + (\lambda(1 - \widetilde{\rho}) - \mu)\exp(-(\lambda - \mu)t)}, \tag{6}
$$

where μ is the rate of becoming non-infectious, and $\widetilde{\rho}$ is the sampled proportion at the present which we estimate separately [83]. This extra factor must be included to avoid double counting of transmission events, as we assume transmission events occur at coalescences within the tree. These internal nodes are the transmission events which result in a descendant which is sampled at the present, and occur with probability $1 - p_0(T_{MRCA} - t)$. We wish to determine the probability of a branch beginning in state i and ending in state j in order to then evaluate the likelihood over the entire tree. If we define $\mathbf{P}(t) = (P_{(0,0)}(t), P_{(0,1)}(t), P_{(1,0)}(t), P_{(1,1)}(t))$ to be the vector of probabilities of observing each state at time t, this amounts to numerically evaluating the solution to

$$
\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{Q}(t) \tag{7}
$$

at the end of each branch within the tree, and subject to a prior on the state at the root, and some modifications to ensure a transmission at each internal node, full details of the algorithm are provided in [24].

Epitope	HLA	HLA prop	with escape		Time to escape (years)		Time to reversion (years)		Replicative capacity	
			HLA +ve	HLA −ve	ODE	MAP	ODE	MAP	rank	actual
GGKKKYKLK	B*08	0.143	0/5	3/34	288	29.6	$>10^6$	180	8	0.925
RLRPGGKKK	A*03	0.224	7/11	4/28	3.36	4.41	4.58	4.14	20	0.812
RLRPGGKKK	A*03	0.224	7/11	4/28	3.36	4.41	4.58	4.14	12	0.894
GSEELRSLY	A*01	0.262	5/10	7/29	9.25	6.66	12.7	11.8	13	0.881
ELRSLYNTV*	B*08	0.143	5/5	33/34	*0.001*	*0.001*	$>10^6$	451	3	1.03
[A]TLYCVHQR	A*11	0.133	0/3	0/36	—	—	—	—	4	1.02
TLYCVHQR	A*11	0.133	3/3	23/36	0.00257	0.163	50.3	$>10^6$	5	0.983
[A]ISPRTLNAW	B*57	0.057	1/7	8/31	26.3	4.85	$>10^6$	$>10^6$	9	0.819
QAISPRTLNAW	A*25	0.042	1/1	16/37	0.00133	*0.001*	120	373	7	0.948
KAFSPEVIPMF	B*57	0.057	0/6	0/30	—	—	—	—	22	0.387
TPQDLNTML	B*42	0.003	0/0	0/57	—	—	—	—	10	0.903
ETINEEAAEW	A*25	0.042	1/3	3/54	8.90	4.04	17.4	8.68	17	0.830
TSTLQEQIGW	B*57	0.057	7/8	5/49	0.610	0.680	8.19	10.1	18	0.924
TSTLQEQIGW	B*57	0.057	4/8	12/49	7.25	4.85	162	$>10^6$	16	0.840
NPPIPVGEIY*	B*35	0.196	7/7	45/50	0.149	*0.001*	215	164	1	1.31
KRWIILGLNK	B*27	0.073	2/7	3/49	12.2	26.0	13.4	$>10^6$	23	0.307
KRWIILGLNK	B*27	0.073	5/7	7/49	1.92	4.85	12.2	$>10^6$	21	0.713
RMYSPTSI	B*52	0.051	2/3	30/54	2.59	1.20	$>10^6$	329	6	0.967
DRFYKTLRA	B*14	0.065	0/6	2/51	460	28.6	$>10^6$	16.7	15	0.847
DRFYKTLRA	B*14	0.065	1/6	1/51	21.5	71.8	7.33	$>10^6$	14	0.858
AEQASQEVKNW*	B*44	0.211	11/13	26/44	2.70	8.06	41.1	135	2	1.24
DCKTILKAL	B*08	0.143	2/8	1/49	13.6	13.4	2.21	1.93	19	0.818
ACQGVGPGHK	A*11	0.133	4/7	13/50	5.62	5.01	23.2	74.1	11	0.897

Table S2: **Fitness costs of epitopes in Gag: SSITT**. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [39], obtained using sequence data from the SSITT cohort [14,34]. Proportions of individuals possessing each HLA type are taken from the HLA factsbook [54]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [24]. Estimates in italics lie on the boundary of our uniform prior. The lack of an estimate occurs when none of the sampled individuals possessed the restricting HLA allele. Putative subtype B negatopes are labelled with asterisks. Amino acids with square brackets lie outside of the epitope. Replicative capacity rank is the ranking of the replicative capacity of the escape mutant (where 1 is the mutant strain with the highest replicative capacity). Replicative capacity actual is the measured scaling of replicative capacity relative to a subtype B lab strain [39].

Table S3: **Fitness costs of epitopes in Gag: HOMER.** MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [39], obtained using sequence data from the HOMER cohort [33]. Proportions of individuals possessing each HLA type are taken from the HLA factsbook [54]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [24]. Estimates in italics lie on the boundary of our uniform prior. The lack of an estimate occurs when none of the sampled individuals possessed the restricting HLA allele. Putative subtype B negatotes are labelled with asterisks. Amino acids with square brackets lie outside of the epitope. Replicative capacity rank is the ranking of the replicative capacity of the escape mutant (where 1 is the mutant strain with the highest replicative capacity). Replicative capacity actual is the measured scaling of replicative capacity relative to a subtype B lab strain in [39].

Epitope	HLA	HLA prop	with escape		Time to escape ($subs\ site^{-1}$)		Time to reversion ($subs\ site^{-1}$)		Replicative capacity	
			HLA +ve	HLA −ve	ODE	MAP	ODE	MAP	rank	actual
GGKKKYKLK	B*08	0.143	2/21	9/140	0.327	0.0657	$>10^4$	0.0318	8	0.925
RLRPGGKKK	A*03	0.224	26/35	28/124	0.00743	0.0115	0.0226	0.0257	20	0.812
RLRPGGKKK	A*03	0.224	26/35	28/124	0.00743	0.0115	0.0226	0.0257	12	0.894
GSEELRSLY	A*01	0.262	18/38	30/122	0.0359	0.0238	0.0495	0.0331	13	0.881
ELRSLYNTV*	B*08	0.143	23/23	149/152	*0.00001*	*0.00001*	$>10^4$	3.70	3	1.03
[A]TLYCVHQR	A*11	0.133	0/23	0/153	—	—	—	—	4	1.02
TLYCVHQR	A*11	0.133	16/19	72/131	0.00707	0.00237	0.256	0.0906	5	0.983
[A]ISPRTLNAW	B*57	0.057	1/6	35/151	0.0722	0.00504	$>10^4$	0.0755	9	0.819
QAISPRTLNAW	A*25	0.042	3/5	28/161	0.0113	0.00229	0.177	0.0906	7	0.948
KAFSPEVIPMF	B*57	0.057	0/10	0/163	—	—	—	—	22	0.387
TPQDLNTML	B*42	0.003	0/0	1/174	—	—	—	—	10	0.903
ETINEEAAEW	A*25	0.042	0/6	7/171	1.37	0.0639	$>10^4$	0.0749	17	0.830
TSTLQEQIGW	B*57	0.057	7/8	7/165	0.00185	0.00189	0.0103	0.0105	18	0.924
TSTLQEQIGW	B*57	0.057	5/8	29/157	0.0104	0.00673	0.102	3270	16	0.840
NPPIPVGEIY*	B*35	0.196	23/23	139/147	*0.00001*	*0.00001*	$>10^4$	0.607	1	1.31
KRWIILGLNK	B*27	0.073	5/14	9/164	0.0266	0.0180	0.0290	0.0235	23	0.307
KRWIILGLNK	B*27	0.073	8/10	18/148	0.00368	0.00381	0.0286	0.0318	21	0.713
RMYSPTSI	B*52	0.051	1/2	30/163	0.0190	0.00506	0.293	0.258	6	0.967
DRFYKTLRA	B*14	0.065	5/12	1/158	0.0176	0.0158	0.00276	0.00260	15	0.847
DRFYKTLRA	B*14	0.065	1/14	2/159	0.195	0.136	0.0392	0.0131	14	0.858
AEQASQEVKNW*	B*44	0.211	25/26	97/126	0.00292	0.0218	0.276	0.332	2	1.24
DCKTILKAL	B*08	0.143	1/21	5/139	0.653	0.103	$>10^4$	0.0169	19	0.818
ACQGVGGPGHK	A*11	0.133	7/21	26/142	0.0565	0.0195	0.167	0.0383	11	0.897

Table S4: **Fitness costs of epitopes in Gag: Bloemfontein.** MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [39], obtained using sequence data from the Bloemfontein cohort [35,36]. Proportions of individuals possessing each HLA type are taken from the HLA factsbook [54]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [24]. Estimates in italics lie on the boundary of our uniform prior. The lack of an estimate occurs when none of the sampled individuals possessed the restricting HLA allele. Putative subtype B negatives are labelled with asterisks. Amino acids with square brackets lie outside of the epitope. Replicative capacity rank is the ranking of the replicative capacity of the escape mutant (where 1 is the mutant strain with highest replicative capacity). Replicative capacity actual is the measured scaling of replicative capacity relative to a subtype B lab strain in [39].

Epitope	HLA	HLA prop	with escape		Time to escape (years)		Time to reversion (years)		Replicative capacity	
			HLA +ve	HLA −ve	ODE	MAP	ODE	MAP	rank	actual
GGKKKYKLK	B*08	0.094	1/13	8/190	101	57.5	60.1	28.6	8	0.925
RLRPGGKKK	A*03	0.125	15/15	150/165	*0.892*	*0.001*	$>10^6$	385	20	0.812
RLRPGGKKK	A*03	0.125	15/15	150/165	*0.892*	*0.001*	$>10^6$	385	12	0.894
GSEELRSLY	A*01	0.095	12/18	60/180	3.99	7.10	40.8	119	13	0.881
ELRSLYNTV*	B*08	0.094	13/13	190/191	*0.001*	*0.001*	$>10^6$	6070	3	1.03
[A]TLYCVHQR	A*11	0.029	0/1	5/201	698	67.4	$>10^6$	$>10^6$	4	1.02
TLYCVHQR	A*11	0.029	1/1	156/196	*0.001*	*0.001*	$>10^6$	186	5	0.983
[A]ISPRTLNAW	B*57	0.078	8/9	39/190	0.639	0.843	14.0	16.4	9	0.819
QAISPRTLNAW	A*25	0.009	0/0	51/198	–	–	–	–	7	0.948
KAFSPEVIPMF	B*57	0.078	2/9	13/191	19.8	67.4	20.9	$>10^6$	22	0.387
TPQDLNTML	B*42	0.099	13/41	17/151	13.4	28.6	20.5	84.1	10	0.903
ETINEEAAEW	A*25	0.009	0/0	2/203	–	–	–	–	17	0.830
TSTLQEQIGW	B*57	0.078	7/9	16/191	1.26	1.54	5.64	6.35	18	0.924
TSTLQEQIGW	B*57	0.078	9/9	187/189	*0.001*	*0.001*	$>10^6$	2750	16	0.840
NPPIPVGEIY*	B*35	0.108	2/3	59/199	3.64	12.2	27.4	2130	1	1.31
KRWIILGLNK	B*27	0.029	1/1	3/201	*0.001*	*0.001*	1.90	2.00	23	0.307
KRWIILGLNK	B*27	0.029	1/1	13/202	*0.001*	26.9	8.61	$>10^6$	21	0.713
RMYSPTSI	B*52	0.033	0/0	199/201	–	–	–	–	6	0.967
DRFYKTLRA	B*14	0.068	4/19	1/183	15.0	15.7	1.41	1.45	15	0.847
DRFYKTLRA	B*14	0.068	2/18	9/181	54.5	33.5	51.7	30.5	14	0.858
AEQASQEVKNW*	B*44	0.112	18/34	54/163	9.48	12.6	68.2	272	2	1.24
DCKTILKAL	B*08	0.094	0/13	3/186	3060	164	$>10^6$	26.0	19	0.818
ACQGVGGPGHK	A*11	0.029	0/1	90/196	31.3	0.618	$>10^6$	86.8	11	0.897

Figure S1: **Scatter-plot of escape proportions in HLA matched and HLA mismatched hosts, coloured by population**. Data points in which no individuals in a population are HLA matched (0/0 escape proportion in HLA matched hosts) were excluded. SSITT, Bloemfontein and HOMER data are shown in red, green and blue respectively. Numbers after the abbreviation of each epitope refers to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA.

	Escape		Reversion	
Study	ρ	p	ρ	p
SSITT	0.862	0.00000102	0.320	0.0909
HOMER	0.794	0.0000245	0.556	0.00670
Bloemfontein	0.753	0.0000993	0.500	0.0147

Table S5: Summary table of the Spearman rank correlation coefficient and associated p-values between rate estimates using the our integrated method and the ODE method.
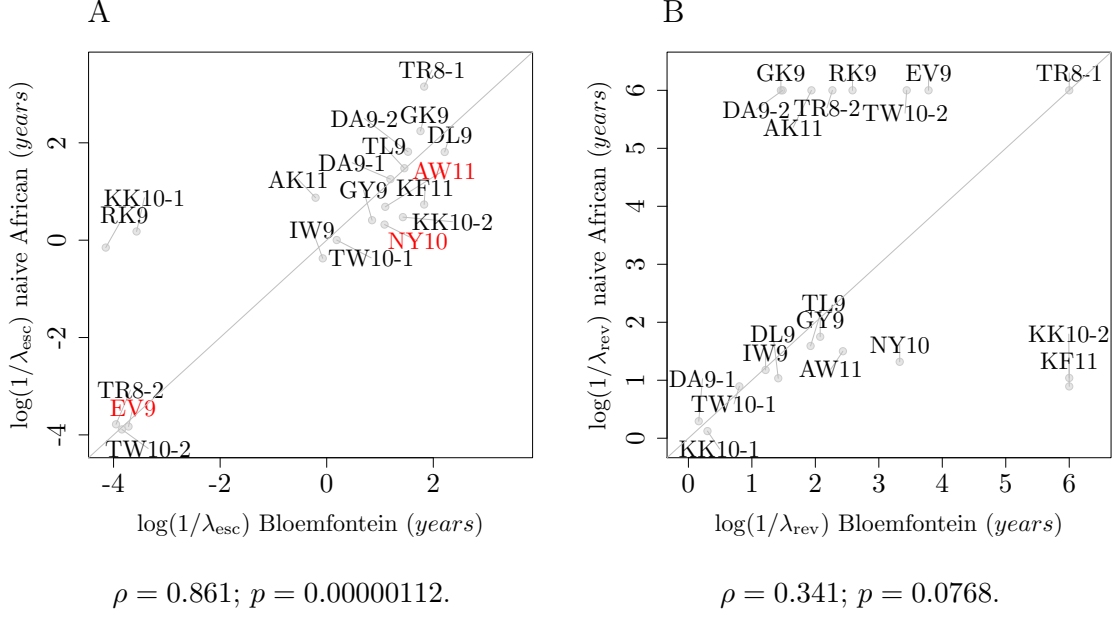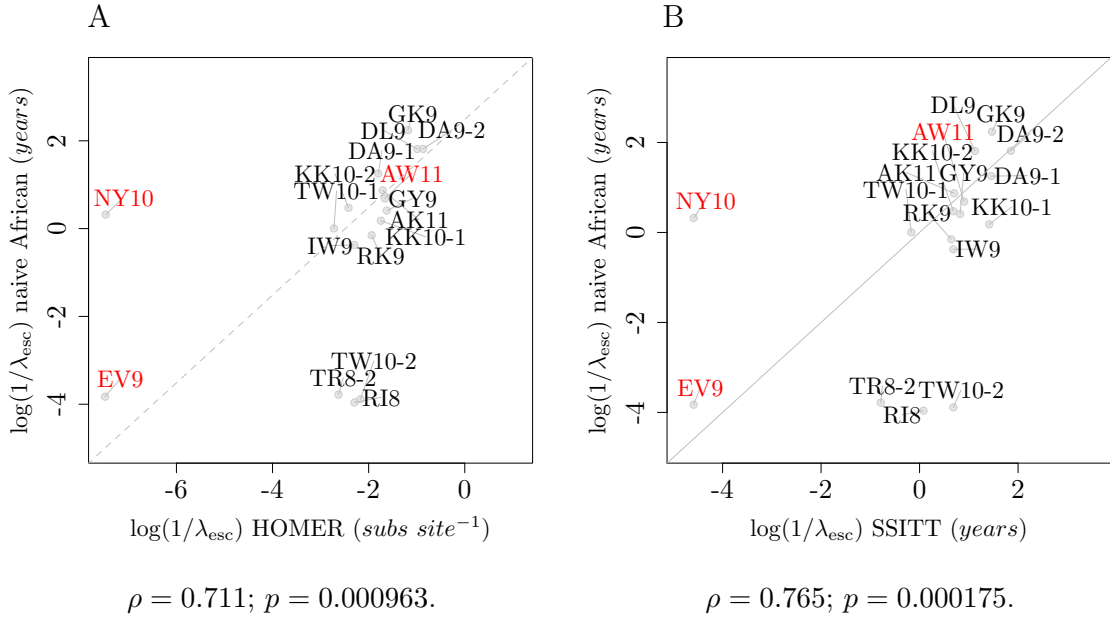
Figure S2: **Comparing time to escape/reversion as estimated from the Bloemfontein data using the integrated method to estimates generated using all available African data and the naive approach**. Spearman rank correlation coefficients (ρ) with associated p-values are displayed. Negatopes are highlighted in red for time to escape. Numbers after the abbreviation of each epitope refers to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. $y = x$ is shown in solid grey.

Figure S3: **Comparing time to escape/reversion as estimated from the SSITT/HOMER data using the integrated method to estimates generated using all available African data and the naive approach**. In Figures S3A - S3B SSITT and HOMER MAP estimates are plotted against a naive estimate of the escape rate. Spearman rank correlation coefficients (ρ) with associated p-values are displayed. Negatopes are highlighted in red. Numbers after the epitope abbreviation refer to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. We also plot $y = x$ in solid grey where estimates are on the same timescale. Dotted grey lines represent an estimate of $y = x$ after a change of timescale assuming the Swiss and Canadian epidemics are expanding at roughly the same rate.

A

B

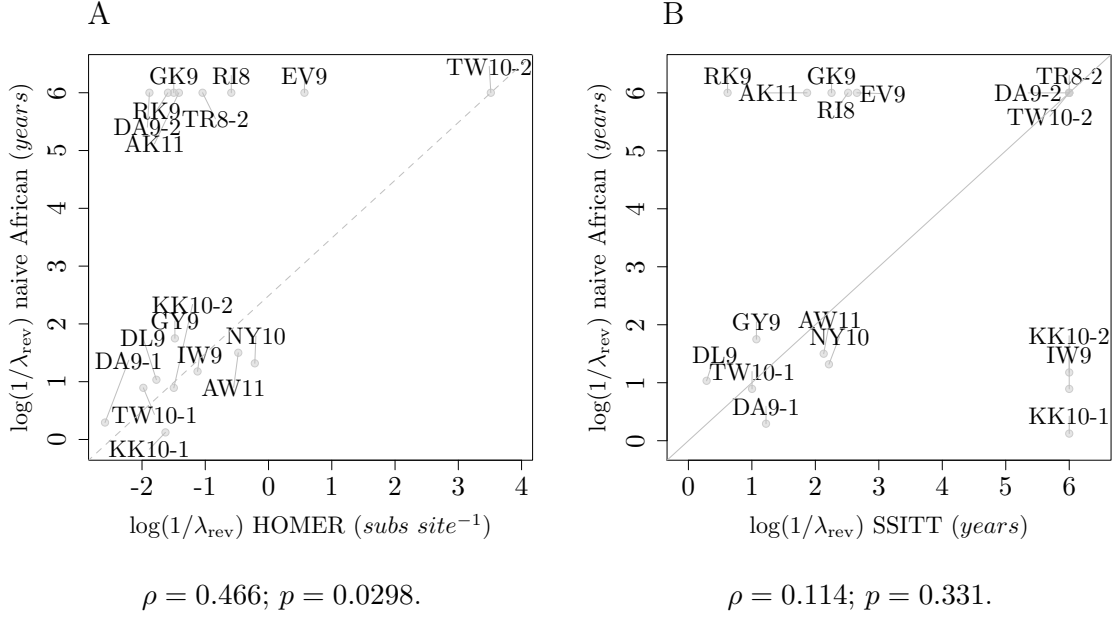$\rho = 0.466; p = 0.0298.$

$\rho = 0.114; p = 0.331.$

Figure S4: **Comparing time to reversion as estimated from the SSITT/HOMER data to estimates of time to reversion using the large African dataset and the naive approach**. In Figures S4A - S4B SSITT and HOMER MAP estimates are plotted against a naive African estimate of the reversion rate. Spearman rank correlation coefficients (ρ) with associated p-values are displayed. Numbers after the abbreviation of each epitope refers to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. $y = x$ is plotted in solid grey where estimates are on the same timescale. Dotted grey lines represent an estimate of $y = x$ after a change of timescale assuming the Swiss and Canadian epidemics are expanding at roughly the same rate.

	Escape		Reversion	
Study	ρ	p	ρ	p
SSITT	0.765	0.000175	0.114	0.331
HOMER	0.711	0.000963	0.466	0.0298
Bloemfontein	0.861	0.00000112	0.341	0.0768

Table S6: Summary table of the Spearman rank correlation coefficient and associated p-values between escape and reversion rate estimates using the ODE model applied to the large African dataset, and the SSITT, HOMER and Bloemfontein datasets.
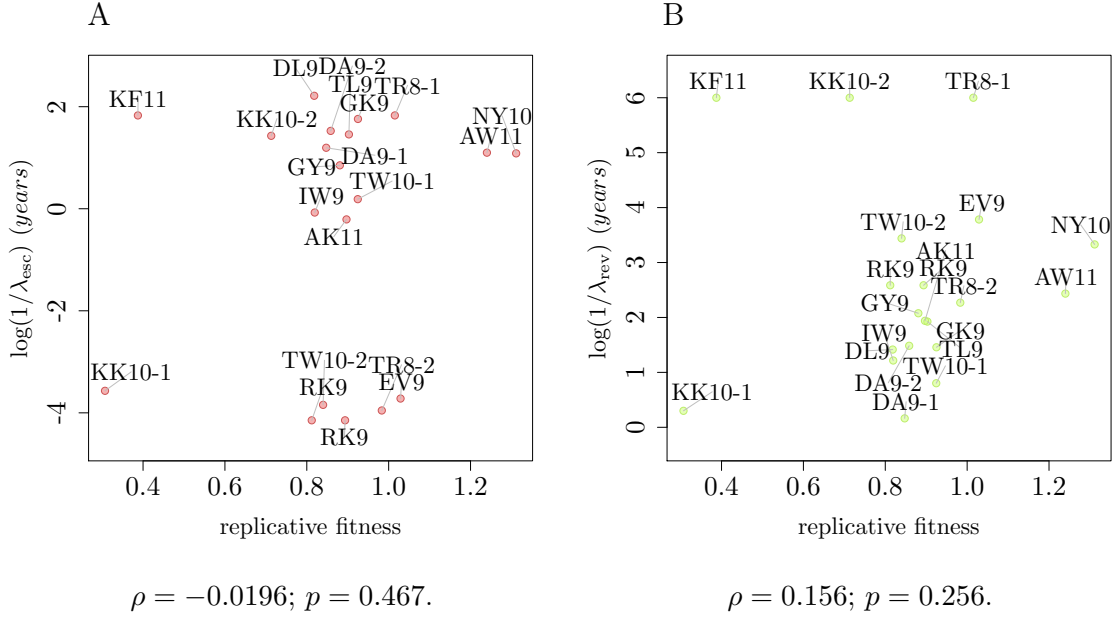
$\rho = -0.0196;\ p = 0.467.$ $\rho = 0.156;\ p = 0.256.$

Figure S5: **Scatter-plots of time to escape and reversion estimates against** *in vitro* **viral replicative capacity: Bloemfontein**. Replicative capacity plotted against the MAP estimates of time to escape is shown in Figure S5A, and the MAP estimates of time to reversion in Figure S5B. Epitopes are labelled by the first three amino acids. Numbers after the abbreviation of each epitope refers to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Spearman rank correlation coefficients; ρ, with associated p-values are displayed on each plot.

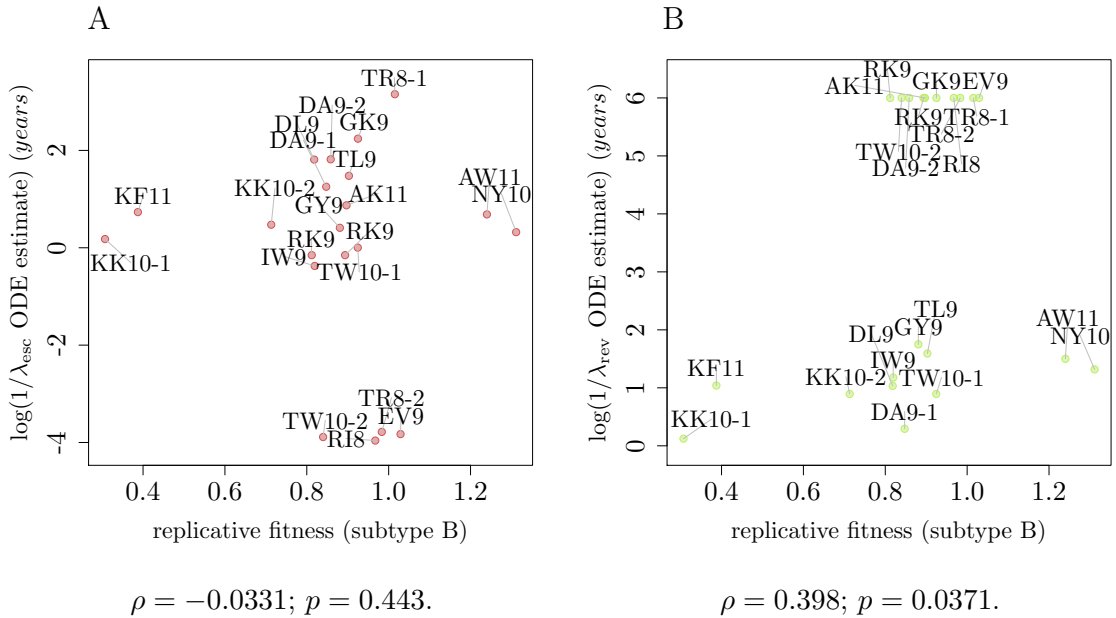$\rho = -0.0331;\ p = 0.443.$ $\rho = 0.398;\ p = 0.0371.$

Figure S6: **Scatter-plots of estimates of time to escape and reversion estimated from all African data using naive ODE approach against** *in vitro* **replicative capacity**. Figure S6A shows a scatter-plot of the respective escape rate estimates. Figure S6B shows the corresponding plot for reversion rate estimates. Numbers after abbreviation of each epitope refers to the ordering in Tables S2 - S4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Spearman rank correlation coefficients with associated p-values are shown.
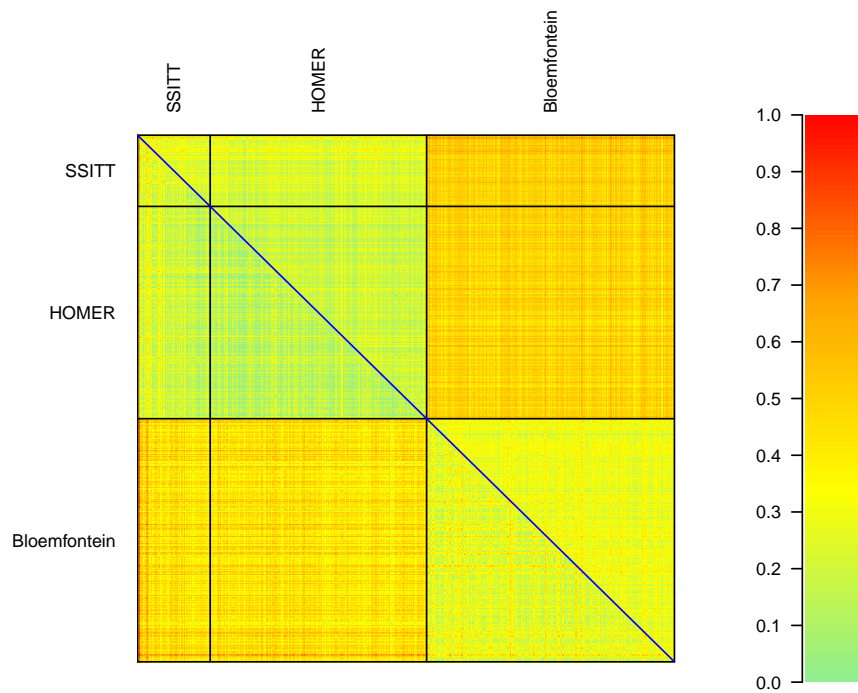
Figure S7: **Sequence divergence by Hamming distance**. We determine the Hamming distance between each pair of sequences in SSITT, HOMER and Bloemfontein for nucleotide and amino acid sequences. We then scale by sequence length to obtain a measure of sequence divergence. Nucleotide sequence divergence is shown above the leading diagonal, and amino acid sequence divergence is plotted below the leading diagonal. Black lines distinguish the SSITT, HOMER, Bloemfontein cohorts. The leading diagonal is shown in blue.
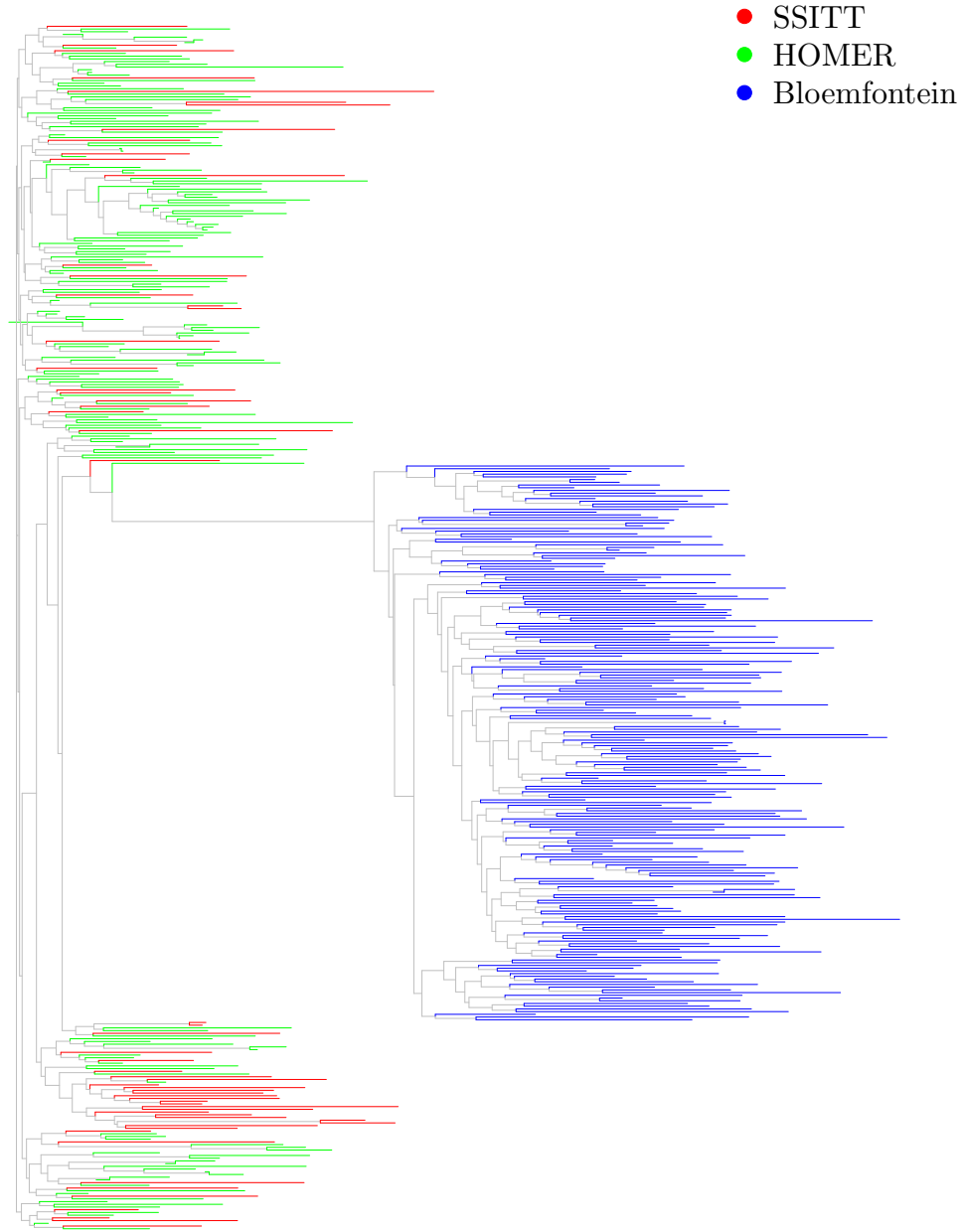
Figure S8: **Neighbour joining tree**. We show a neighbour joining tree obtained using the K81 [68] model applied to the largest region of *gag* over which $< 10\%$ sequences contained a gap or an unknown nucleotide in the sequence alignment. Terminal branches are coloured by cohort according to the legend.